

ОГЛЯД ПОТОЧНОГО СТАНУ ПИТАННЯ ВИЯВЛЕННЯ ГЕНЕРАЦІЇ ТЕКСТУ ШТУЧНИМ ІНТЕЛЕКТОМ

Герасимова В. А.

e-mail: vladyslava.herasymova@nure.ua

Науковий керівник – к.т.н., доц. Мельнікова Р. В.

Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

This paper outlines the current state of the field of AI-generated text detection by providing an overview of the theoretical basis behind it and an analysis of the publicly available AI text detection tools. To do the later, two datasets of human-written and AI-generated text, one in English and one in Ukrainian, were formed and used to examine four different publicly available AI text detection tools, namely ZeroGPT, Reilaa, Surfer and AiDetector.com. The study concludes that, while texts in English might be analyzed with some degree of caution, the analysis of texts in Ukrainian for now is too unreliable.

Можливість аналізу тексту та виявлення того, чи було його згенеровано штучним інтелектом, підлягає все більш активним дискусіям кожного року через все швидший розвиток асистентів генеративного штучного інтелекту і, відповідно, все більше семантичне наближення тексту, згенерованого ними, до людської мови, що ускладнює розв'язання цієї проблеми [1, 2].

На базі цього було вирішено провести дослідження, задачею якого є огляд наявної теоретичної бази питання виявлення генерації тексту штучним інтелектом, а також проведення практичного дослідження наявних на поточний момент у вільному доступі сервісів по розв'язку цієї задачі.

Потенційні та використовувані наразі методики виявлення генерації тексту штучним інтелектом можна класифікувати на 5 типів [3]:

1. Із використанням водяних знаків різних типів.
2. Методи на основі ознак.
3. Нейронні методи.
4. Методи, що використовують допомогу людини.
5. Гібридні методи.

Оскільки головним фокусом дослідження є саме автоматизований аналіз тексту, а також враховуючи той факт що гібридні методи представляють собою комбінацію інших, більше уваги було приділено саме першим трьом типам методик.

Використання водяних знаків передбачає додавання певних додаткових метаданих, а саме інформації щодо того, що текст було згенеровано певним генеративним ШІ, безпосередньо в тексті таким чином, що текст не змінюється достатньо суттєво, щоб цю водяну марку можна було б помітити.

Методи аналізу текстів на основі ознак використовують вимірювану різницю між такими характеристиками текстів як синтаксис, граматики, емоціональність, тощо. Основними такими методами є стилістичний аналіз, аналіз на базі частотних ознак, а також аналіз на базі таких статистичних метрик як розгубленість, складність речення, ентропія та рівномірна щільність інформації [3].

Нейронні методи аналізу текстів передбачають використання попередньо натренованих мовних моделей, що дозволяє пропустити крок «ручного» виявлення значень статистичних характеристик передаючи весь текст на вхід та очікуючи конкретну відповідь щодо вірогідності його генерації ШІ на вихід. Виявлення значимих статистичних характеристик та їх очікуваних значень відбувається під час тренування моделі [2].

Протягом практичної частини дослідження було розглянуто 11 різних програмних систем, для яких виявлення штучної генерації тексту є головною чи додатковою функціональністю. Серед них було обрано ті, які не мають обмежень у кількості використань без реєстрації, а також тестування яких може бути оптимізовано через відправлення запитів за допомогою Python скрипту: ZeroGPT, Reilaa, Surfer та AiDetector.com.

Для аналізу обраних програмних систем було сформовано два датасети: один англійською мовою, і другий українською мовою. В кожному з них є 100 текстів написаних людиною (новинні та інформаційні статті опубліковані до 2022 року та старі (до 2022 року) версії статей з вікіпедії) та 100 текстів згенерованих за допомогою таких ШІ моделей як ChatGPT, Claude та Gemini. В результаті було отримано 5 наборів даних (сервіс Reilaa повертає 2 відповіді) із вірогідностями штучної генерації відповідних текстів, подальший аналіз яких проводився за допомогою граничного значення. Додаткове дослідження продемонструвало що в залежності від конкретної кінцевої мети дослідження (максимізація точності, мінімізація рівню хибно позитивних результатів, тощо) та допустимих значень інших характеристик це критичне значення може різнитися.

Результати аналізу результатів для граничного значення 50% для текстів англійською та українською мовами наведено на рис. 1, 2. На них відображено значення таких кількісних результатів як кількість істинно позитивних, істинно негативних, хибно позитивних та хибно негативних результатів, а також такі статистичні характеристики як точність, відгук, рівень хибно позитивних результатів та влучність.

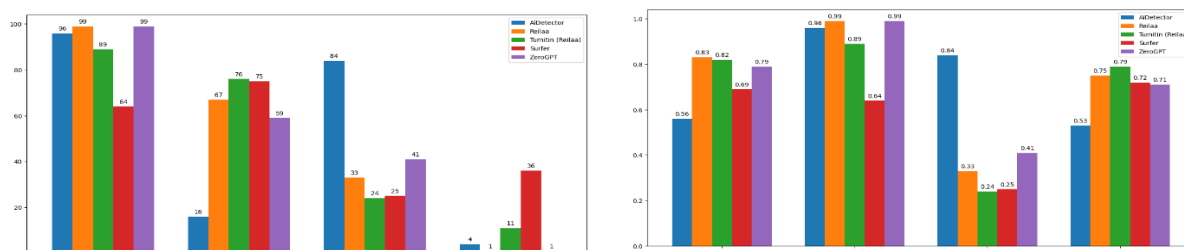


Рисунок 1 – результати аналізу текстів англійською мовою

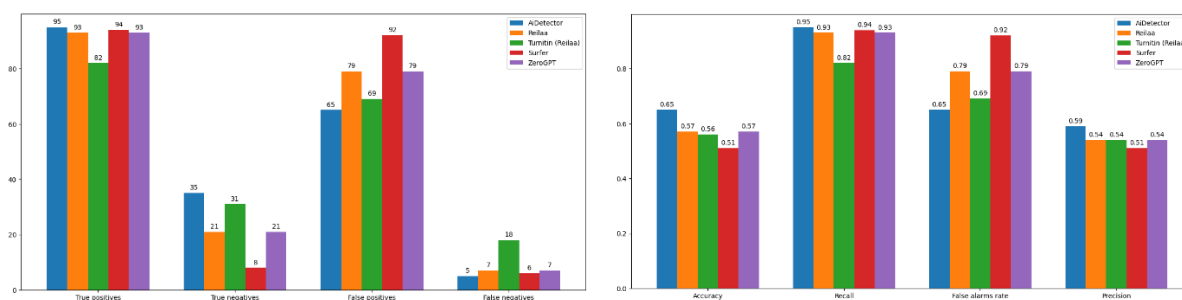


Рисунок 2 – результати аналізу текстів українською мовою

На базі цих результатів було зроблено декілька висновків, а саме:

1. Переважна більшість моделей при аналізі текстів як англійською, так і українською мовою, коректно розпізнає більшість текстів, згенерованих штучним інтелектом;
2. Майже всі моделі продемонстрували гірші результати розпізнавання текстів, написаних людиною, ніж текстів, згенерованих ШІ, обома мовами;
3. При аналізі текстів українською мовою написаних людиною всі моделі зробили більше помилкових ніж коректних рішень та продемонстрували достатньо низькі значення влучності, що свідчить про недоцільність використання розглянутих сервісів для аналізу текстів українською мовою;
4. При аналізі текстів англійською мовою написаних людиною було отримано значно нижчі, але все ще достатньо високі значення рівня хибно позитивних результатів, що свідчить про те що використання розглянутих сервісів для аналізу текстів англійською мовою наразі хоча і є більш можливим ніж текстів українською мовою, але все одно потребує обережності.

Список використаних джерел:

1. Chakraborty S., Bedi A. S., Zhu S., An B., Manocha D., Huang F. On the Possibilities of AI-Generated Text Detection. DOI: <https://doi.org/10.48550/arXiv.2304.04736>
2. Fraser K. C., Darkins H., Kiritchenko S. Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. Journal of Artificial Intelligence Research. 2025. Vol. 82. P. 2233–2278 DOI: <https://doi.org/10.1613/jair.1.16665>
3. Fariello S., Fenza G., Forte F., Gallo M., Marotta M. Distinguishing Human From Machine: A Review of Advances and Challenges in AI-Generated Text Detection. International Journal of Interactive Multimedia and Artificial Intelligence. 2025. Vol. 9, No3. P. 6–18. DOI: <https://doi.org/10.9781/ijimai.2024.12.002>